

ABSTRACT OF THE DISCLOSURE

There is disclosed a document type definition
generating method comprising, in a structured document
provided with a tag having an element name in each
5 document element, judging a physical structure of each
document element from indentation, blank lines, and
positional relation between tags, analyzing words and
phrases in each document element, and judging a
semantic structure of the document element based on
10 words and phrases connection and word types. When the
physical and semantic structures of document elements
having tags different in element name are similar, the
elements are regarded as being of the same type and one
element name is excluded from a list for generating the
15 document type definition. When the physical and
semantic structures of document elements having tags
with the same element name are different, the elements
are regarded as being of the different types and one
element name is changed. Furthermore, the words and
20 phrases between a start tag and an end tag with the
same title are analyzed, and the information to be
included between the tags is obtained to generate the
document type definition. Thereby, tag meaning is
correctly treated, and the document type definition
25 with tag redundancy removed therefrom is generated.

09449706.112499